

# OWL-Based Validation of Biomodels

## a project in Semantic Systems Biology



**Michel Dumontier, Ph.D.**

**Associate Professor of Bioinformatics, Department of Biology, School of Computer Science, Institute of Biochemistry, Carleton University**

**Professeur Associé, Université Laval**

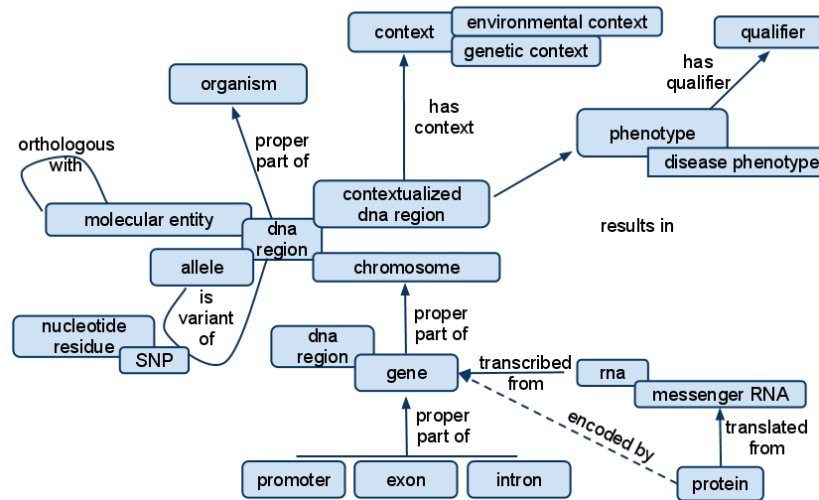
**Ottawa Institute of Systems Biology  
Ottawa-Carleton Institute of Biomedical Engineering**

# Semantic Systems Biology

An interdisciplinary approach that combines computational biology with formal semantics in order to accurately build, investigate and validate quantitative models of biological phenomena.

- michel dumontier

# Bio-ontologies



- Provide rich **human and machine understandable descriptions** of the common terminology
- Generally used for **semantic annotation** of data, which when reused, facilitate integration, search and retrieval
- Facilitate **granular and cross-domain queries** on their contents
- Automated reasoning over formal ontologies can be used to generate **explanations for entailments**

# SBML-based biomodels place semantic annotations in an annotation element

subject and attributes

→ `<species metaid="_525530" id="GLCi" compartment="cyto" initialConcentration="0.097652231064563">`



*The annotation element stores the RDF annotation*

`<annotation>  
 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/" xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0#" xmlns:bqbiol="http://biomodels.net/biology-qualifiers/" xmlns:bqmodel="http://biomodels.net/model-qualifiers/"`

subject



`<rdf:Description rdf:about="#_525530">`

predicate  
(qualifier)

→ `<bqbiol:is>  
 <rdf:Bag>  
 <rdf:li  
 rdf:resource="urn:miriam:obo.chebi:CHEBI%3A4167"/>  
 <rdf:li rdf:resource="urn:miriam:kegg.compound:C00031"/>  
 </rdf:Bag>  
</bqbiol:is>  
</rdf:Description>  
</rdf:RDF>  
</annotation>  
</species>`

object



The *intent* is to express that the species represents a substance composed of glucose molecules  
We also know from the SBML model that this substance is located in the cytosol and with a (initial) concentration of 0.09765M

## By converting models into formal representations of knowledge we get to:

- **capture** the semantics of models and the biological systems they represent
- **leverage** knowledge explicit in linked terminologies
- **validate** the accuracy of the annotations / models
- **discover** biological implications inherent in the models
- **query** the results of simulations in the context of the biological knowledge

# Have you heard of OWL?



# OWL - Web Ontology Language – enables more accurate knowledge capture than RDF

**Enhanced vocabulary (strong axioms)** to express knowledge relating to classes, properties, individuals and data values

- quantifiers
  - existential, universal, cardinality restriction
- negation
- disjunction
- property characteristics
  - transitive, functional, inverse functional, symmetric, antisymmetric, reflexive, irreflexive
- complex classes in domain and range restrictions
- property chains

# Reasoning over OWL ontologies

- **Consistency:** determines whether the ontology contains contradictions.
- **Satisfiability:** determines whether classes can have instances.
- **Subsumption:** is C2 implicitly a subclass of C1?
- **Classification:** repetitive application of subsumption to discover implicit subclass links between named classes
- **Realization:** find the most specific class that an individual belongs to.



# Triples to axioms: Many possible formalizations – knowledge of logics and domain expertise comes in handy here!

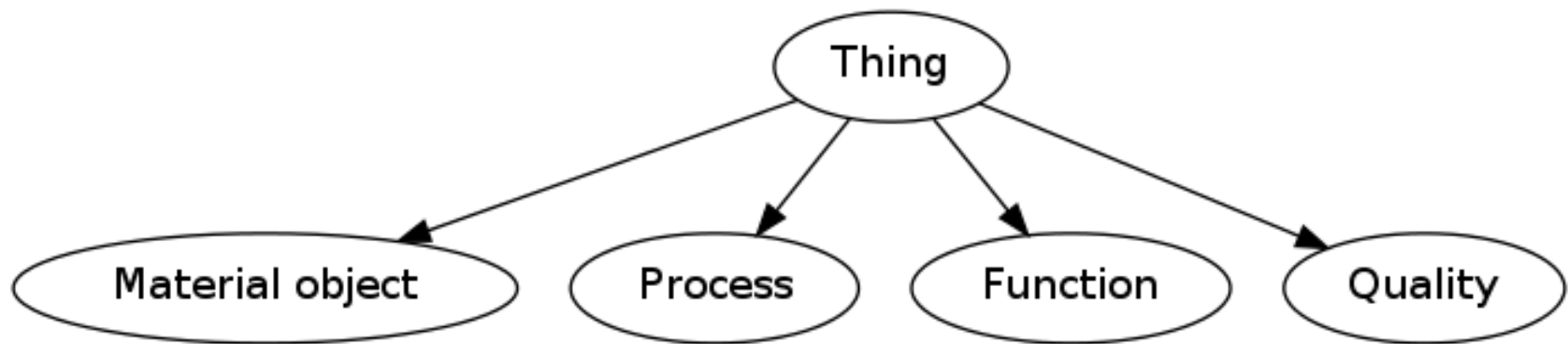
We make an *ontological commitment* by converting RDF triples into OWL axioms that have a specific meaning

Triple in RDF:

<C1 R C2>

- If C1 and C2 are classes, and R a relation between 2 classes
- specify meaning:
  - C1 SubClassOf: C2
  - C1 SubClassOf: R some C2
  - C1 SubClassOf: R only C2
  - C2 SubClassOf: R some C1
  - C1 SubClassOf: S some C2
  - C1 DisjointFrom: C2
  - C1 and C2 SubClassOf: owl:Nothing
  - R some C1 DisjointFrom: R some C2
  - C1 EquivalentClasses C2
  - ...
- in general:  $P(C1, C2)$ , where P is an OWL axiom (template)

# Top-level ontologies can make additional commitment by enforcing *disjointness* among basic types



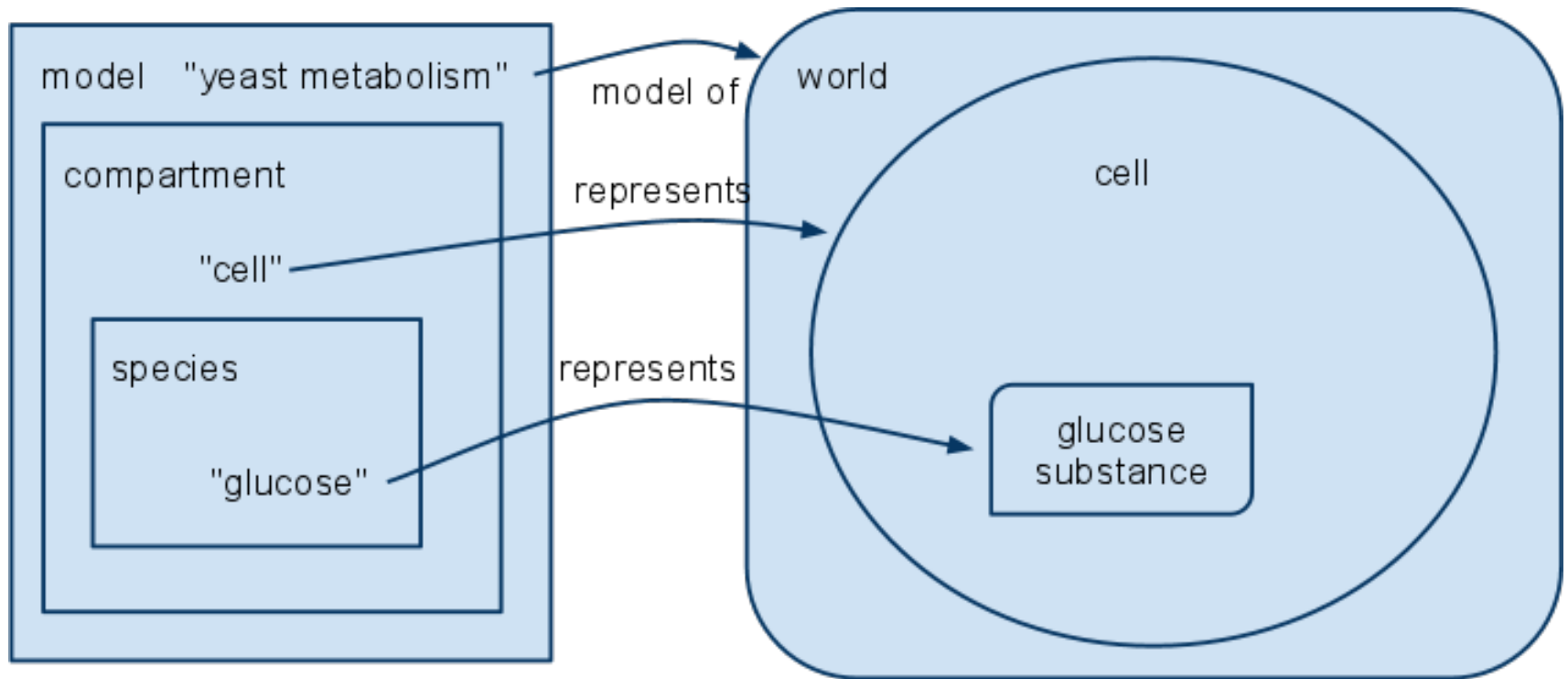
Material object, Process, Function and Quality are mutually disjoint.

# Relations impose additional constraints, such that inconsistencies arise when incorrectly used

Relation	Domain	Range	Inverse
part-of participates-in function-of realizes occurs-in quality-of input-of output-of modifier-of	Entity Individual Function Process Process Quality Physical object Physical object Physical object	Entity Process Physical object Function Physical object Individual Process Process Process	has-part has-participant has-function realized-by has-process-occurring has-quality has-input has-output has-modifier
represents model-of	Model entity Model	Physical object Physical object	

## Conceptualization:

Models and their components represent physical entities (material entities, processes)



## Formalization:

every element  $E$  of the SBML language represents a class  $\text{Rep}(E)$  and we assert that

**$E$  subClassOf: represents some  $\text{Rep}(E)$**

# For each model annotation, we make a commitment to what it represents

OWL Axiom:

M **SubClassOf**: represents some MaterialEntity

Conversion rule: a Model annotated with class C represents:

If C is a SubClassOf **MaterialEntity** then

M **SubClassOf**: represents some C

If C is a SubClassOf **Function** then

M **SubClassOf**: represents some (has-function some C)

If C is a SubClassOf **Process** then

M **SubClassOf**: represents some (has-function some (realized-by only C))

# BIOMODEL 82: Converting *Model*

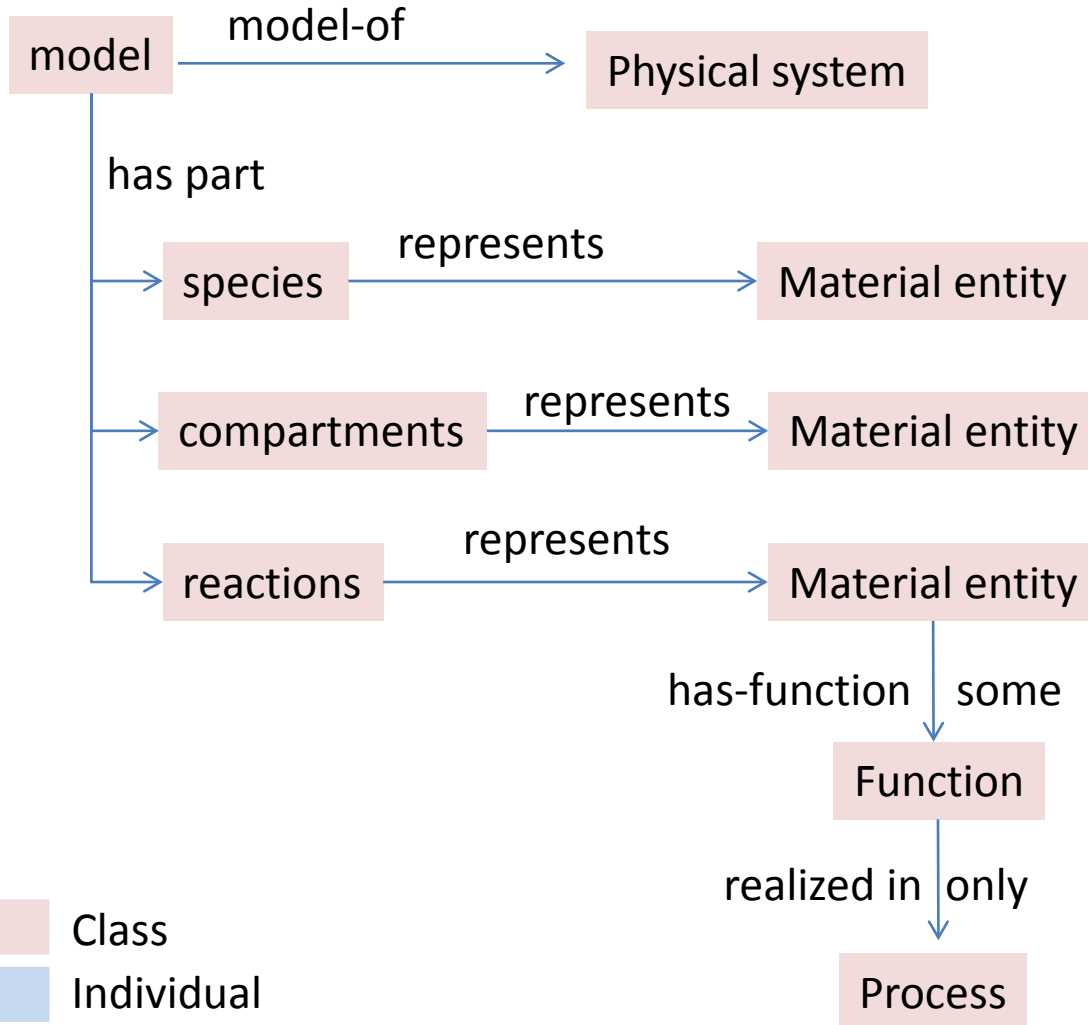
Annotated with **heterotrimeric G-protein complex cycle (GO:0031684)**:

*represents* an object O that has the capability F that is realized in processes of the type *heterotrimeric G-protein complex cycle*

M SubClassOf: represents some O1

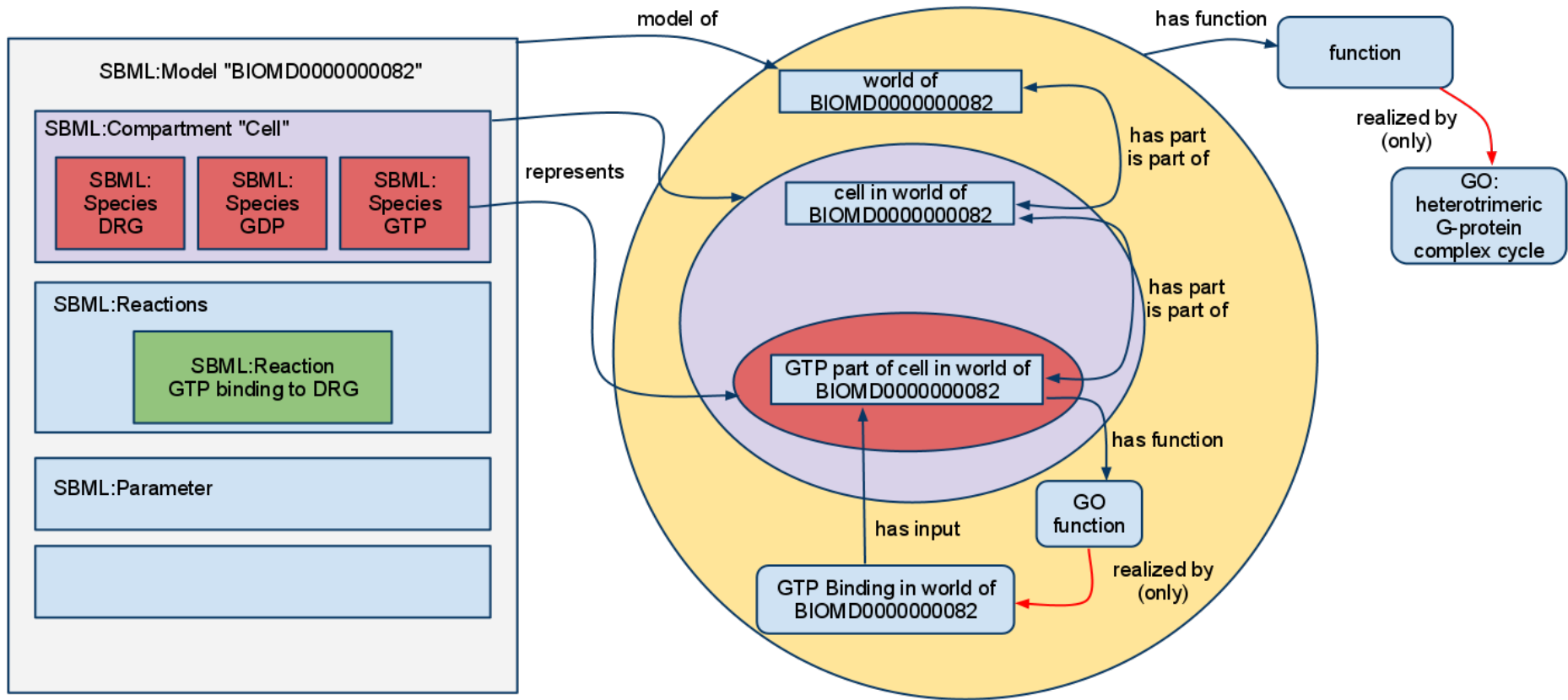
O1 SubClassOf: (has-function some (realized-by only GO:0031684))

# Models and their components represent physical entities (material entities, processes)



## SBMLHarvester

Robert Hoehndorf, Michel Dumontier, John H Gennari, Sarala Wimalaratne, Bernard de Bono, Daniel L Cook and Georgios V Gkoutos. [Integrating systems biology models and biomedical ontologies](#). BMC Systems Biology 2011, 5:124.





# SBML2OWL: Implementation

SBML Harvester <http://code.google.com/p/sbmlharvester/>

- libSBML to access model structure & extract RDF annotations
- Jena RDF API to parse RDF annotations
- OWLAPI to create OWL axioms & reason with top-level ontology

Application to BioModels repository yields:

- OWL ontology with more than 300,000 classes, 800,000 axioms
- includes all referenced ontologies
  - GO (functions, compartments, processes)
  - ChEBI (molecules)
  - Celltype (cell types)
  - FMA (anatomy)
  - PATO (qualities)

# Answering questions

Query	Query string	# results
Contradictory defined entities	Nothing	4,899
Models which represent a process involving sugar	<code>model-of some (has-part some (has-function some (realized-by only (has-participant some sugar))))</code>	54
Parts of BIOMD00000000015 that represent processes involving sugar	<code>part-of some BIOMD00000000015 and represents some (has-function some (realized-by only (has-participant some sugar)))</code>	29
Model entities that represent the cell cycle	<code>represents some (has-part some (has-function some (realized-by only 'cell cycle')))</code>	14
Model entities that represent mutagenic central nervous system drugs in the gastrointestinal systems	<code>represents some (has-part some ('has role' some 'central nervous system drug' and 'has role' some mutagen and part-of some 'Gastrointestinal system'))</code>	2
Model entities that represent catalytic activity involving sugar in the endocrine pancreas	<code>represents some (has-function some (realized-by only (realizes some 'catalytic activity' and has-participant some (sugar and contained-in some (part-of some 'Endocrine pancreas')))))</code>	4

# Model verification

After reasoning, we found 27 models to be inconsistent

*reasons*

1. **our representation** - functions sometimes found in the place of physical entities (e.g. entities that secrete insulin). better to constrain with appropriate relations
2. **SBML abused** – e.g. species used as a measure of time
3. **Incorrect annotations** - constraints in the ontologies themselves mean that the annotation is simply not possible

# Finding inconsistencies with axiomatically enhanced ontologies

related work treated function as a process and described ATPase activity (GO:0004002) as a Catalytic activity that has Water and ATP as input, ADP and phosphate as output and is a part of an ATP catabolic process.

To this, we add:

- GO: ATP + Water the only inputs (universal quantification)
- ChEBI: Water, ATP, alpha-D-glucose 6-phosphate are all different (disjointness)

BIOMD00000000176 and BIOMD00000000177 models of anaerobic glycolysis in yeast.

- “ATP” input to “ATPase” reaction, which is annotated with ATPase activity. The species “ATP”, however, is mis-annotated with Alpha-D-glucose 6-phosphate (CHEBI:17665), not with ATP.

# Species are further described with 'modifiers' *in the context of a reaction*

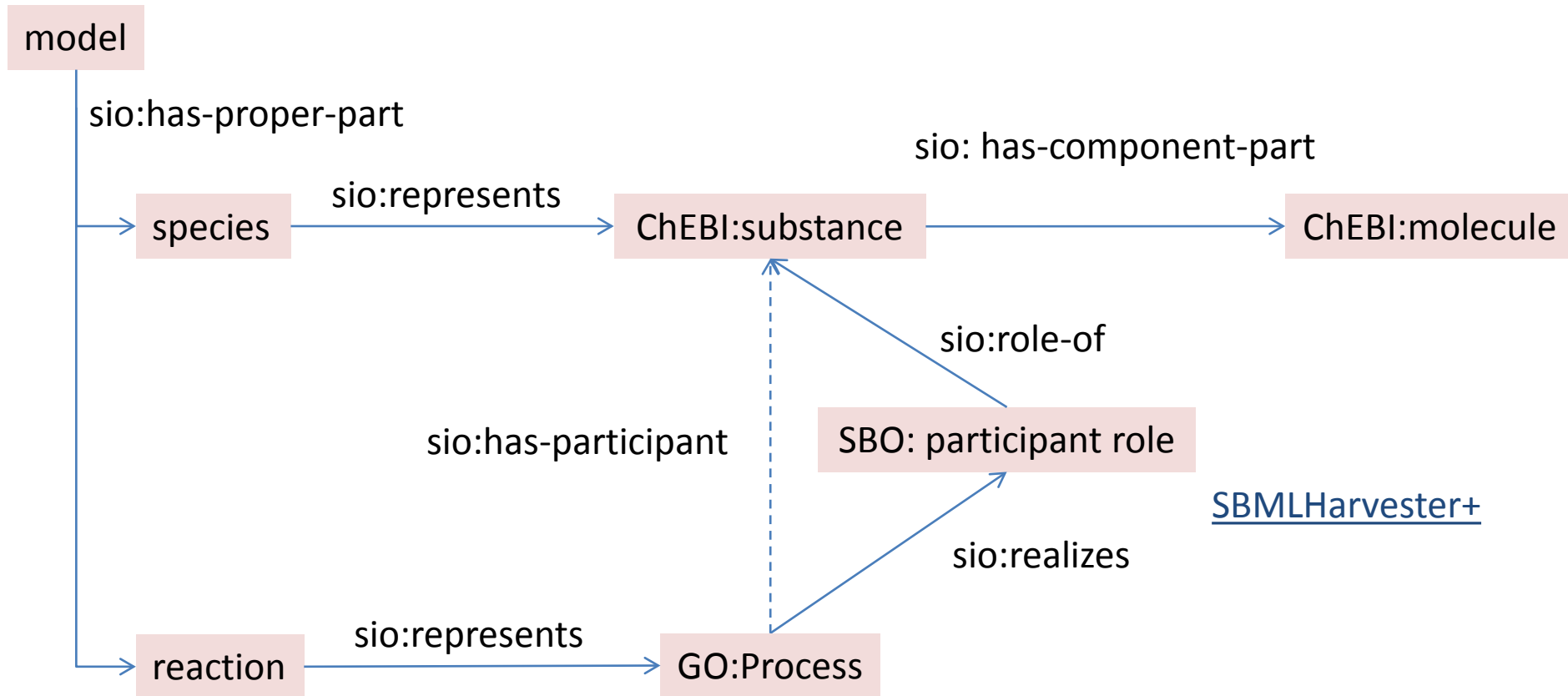
## **essential activator**

```
<listOfModifiers>  
  <modifierSpeciesReference sboTerm="SBO:0000461" species="X"/>  
</listOfModifiers>
```

## **partial inhibitor**

```
<listOfModifiers>  
  <modifierSpeciesReference sboTerm="SBO:0000536" species="PX"/>  
</listOfModifiers>
```

# Roles are realized in the context of processes



Class

Individual

Datatype

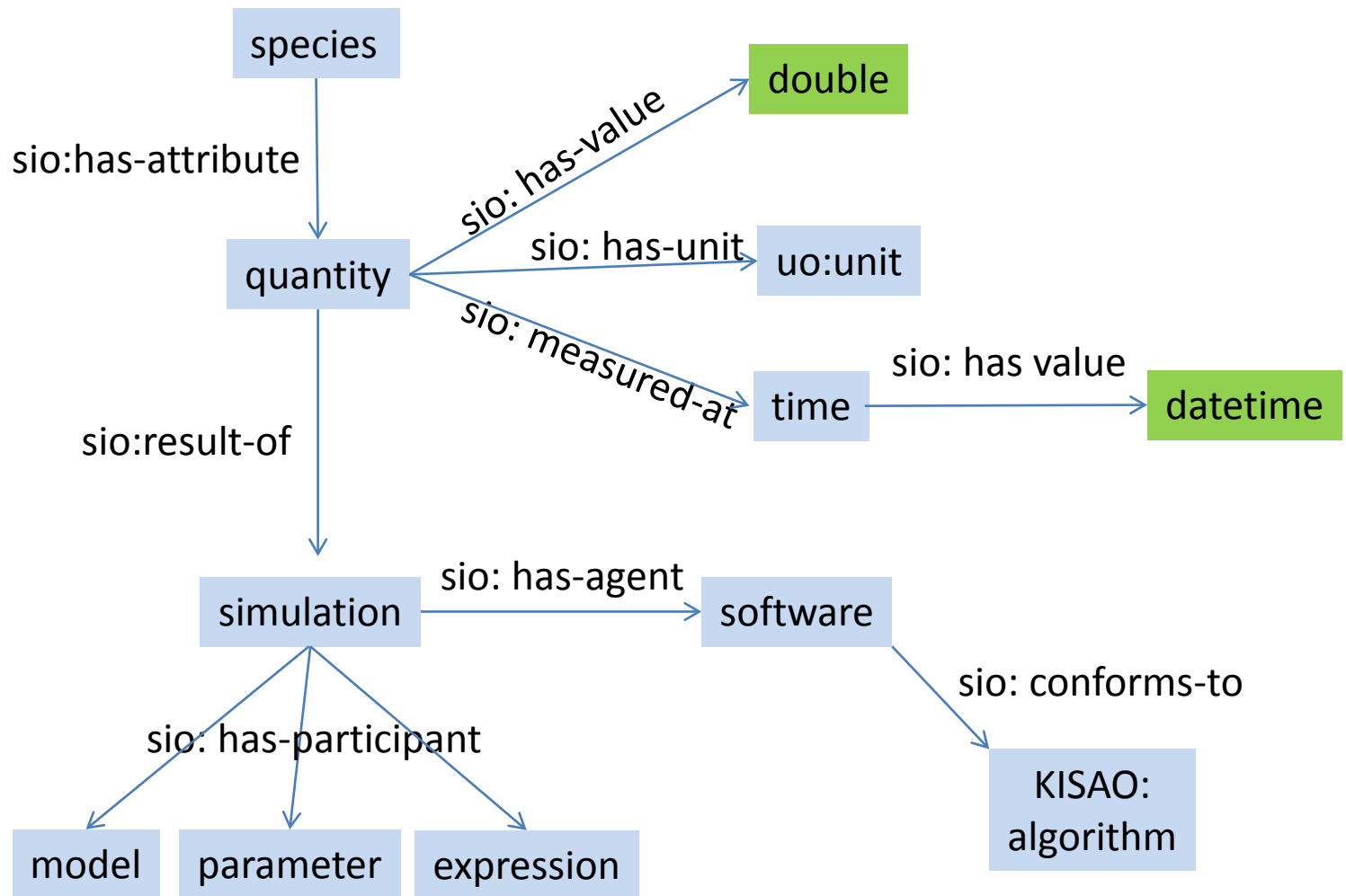
*role chain*: realizes o role of -> has participant

sio: <http://semanticscience.org/resource/>

# Semanticscience Integrated Ontology (SIO)

- OWL2 ontology
- 1000+ classes covering basic types (physical, processual, abstract, informational) with an emphasis on biological entities
- 184 basic relations (mereological, participatory, attribute/quality, spatial, temporal and representational)
- axioms can be used by reasoners to compute inferences for consistency checking, classification and answering questions about life science knowledge
- embodies emerging ontology design patterns
  - *specifies a data model*
- dereferenceable URIs – used in Bio2RDF and SADI
- searchable in the NCBO bioportal
- Available at <http://semanticscience.org/ontology/sio.owl>

# When running a simulation, certain quantities can vary with time

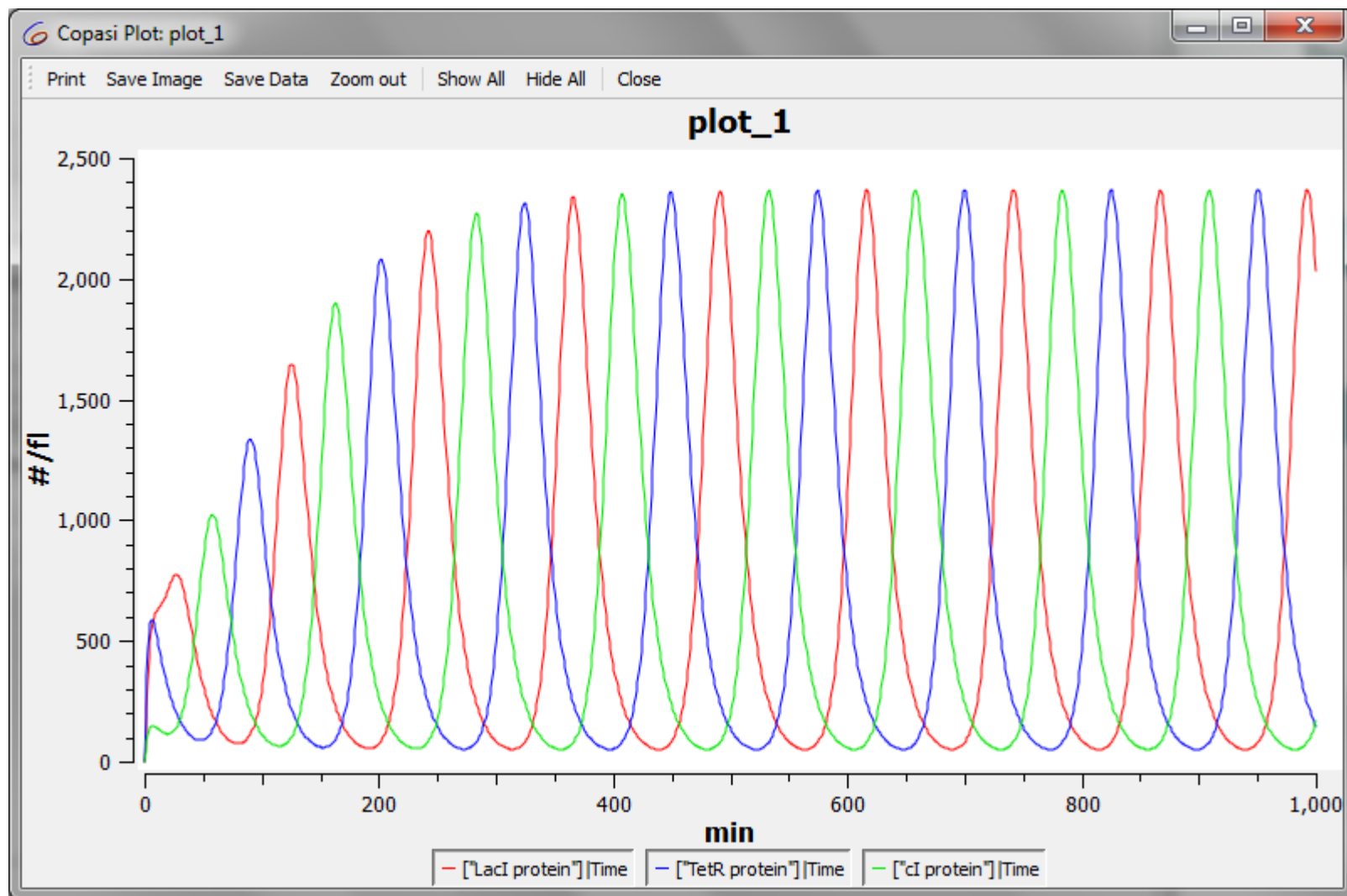


Class

Individual

Datatype





# Copasi output:

## not machine understandable

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
# Time	LacI prote	TetR prote	cl protein	LacI mRNA	TetR mRN	cl mRNA	Compartm	Values[be	Values[al	Values[al	Values[tr	Values[n	Values[K	Values[m	Values[pr	Values[av
0	0	0	0	0	20	0	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
1	81.4405	188.382	42.6413	19.9034	30.6156	7.491	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
2	218.539	358.027	84.5888	21.2335	23.6089	5.86627	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
3	337.623	469.627	113.055	18.5441	17.2349	4.40364	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
4	428.939	536.664	131.035	15.5598	12.4821	3.29582	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
5	495.497	572.019	141.441	13.018	9.03773	2.48567	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
6	542.466	585.398	146.532	11.0337	6.56428	1.90266	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
7	574.778	583.832	147.977	9.56151	4.79541	1.48837	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
8	596.57	572.335	146.993	8.51888	3.53331	1.19835	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
9	611.14	554.454	144.458	7.82324	2.63412	0.999828	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
10	621.044	532.674	141.004	7.40287	1.99408	0.869004	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
11	628.231	508.728	137.085	7.19849	1.5387	0.788821	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
12	634.16	483.805	133.028	7.16163	1.21462	0.747164	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
13	639.897	458.714	129.068	7.25233	0.983648	0.735535	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
14	646.193	433.993	125.381	7.43676	0.818491	0.748075	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
15	653.535	409.987	122.095	7.68527	0.699679	0.780855	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
16	662.19	386.909	119.309	7.97095	0.61336	0.83136	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
17	672.232	364.881	117.104	8.26865	0.549728	0.89813	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
18	683.569	343.958	115.547	8.55459	0.501895	0.980494	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
19	695.958	324.155	114.698	8.80638	0.465075	1.07839	1	0.2	0.216404	216.188	20	2	40	2	10	2.88539
--	----	----	----	----	----	----	-	--	-----	-----	--	-	--	-	--	-----

# Query Answering over RDF/OWL

Find those concentration measurements for species that represent molecular entities that contain ribonucleotide residues

'concentration'

and ('measured at' some double[>20.0, <40.0])


and 'is attribute of' some (

'species'

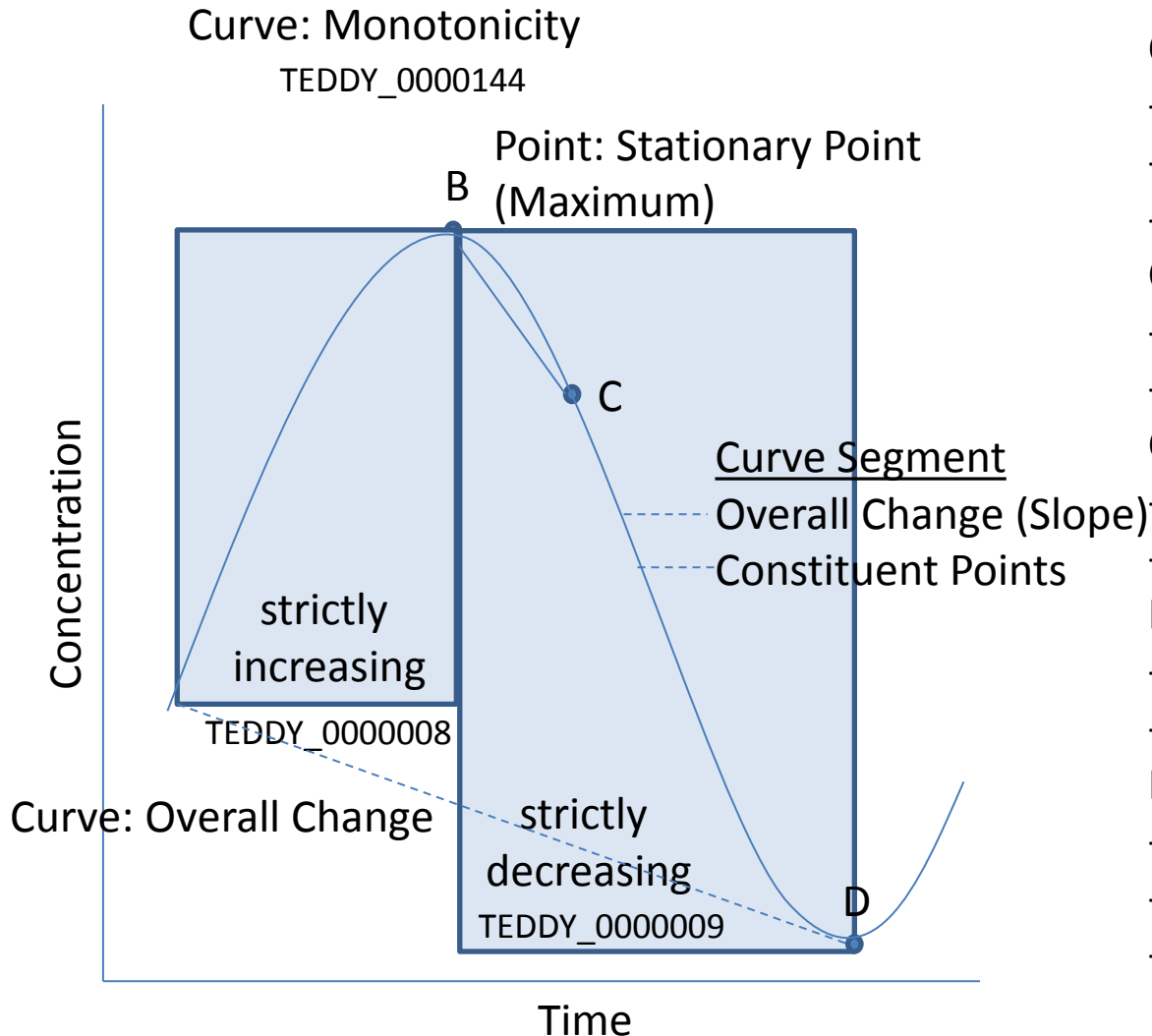
and 'represents' some ('has part' some 'ribonucleotide residue')

)

ChEBI ontology



# Curve Analysis: Elements of a Plot



Curves Contain

- Points
- Line Segments
- Curve Segments

Curves Annotated As

- Monotonic/Non-Monotonic
- Overall Increasing/Decreasing

Curve Segments

- Contain line segments
- Same properties as curve

Line Segments

- Contain points
- Increasing/Decreasing

Points

- Have attributes and values
- Can be local minima/maxima
- Can be inflection points

# Get non-monotonic behaviour for entities that bind to DNA

‘non-monotonic curve’

curve analyzer +  
TEDDY

and ‘has part’ some (

SIO

‘concentration’

and ‘is attribute of’ some (

‘species’

Biomodel +  
UniProt + GO

and ‘has function’ some ‘dna binding’)

)

# Conclusion

The SBML-derived ontologies can be

- i) checked for their consistency, thereby uncovering erroneous curation
- ii) infer attributes and relations of the substances, compartments and reactions beyond what was originally described in the models
- iii) answer sophisticated questions across a model knowledge base
- iv) extended with modifiers, mathematical expressions and parameters, simulation Results (from tab files) to answer questions about simulation results with reference to the semantic annotations (GO) in biomodels, UniProt

# Acknowledgements

Leonid Chepelev

Robert Hoehndorf, John Gennari, Sarah Wimalaratne,  
Bernard de Bono, Daniel Cook, and George Gkoutos.  
Integrating systems biology models and biomedical  
ontologies. BMC Systems Biology. 2011. 5 : 124



**NSERC  
CRSNG**

Canada Foundation for Innovation  
Fondation canadienne pour l'innovation



**Carleton**  
UNIVERSITY



Health  
Canada



**Ontario**



Michel Dumontier

michel\_dumontier@carleton.ca

*Publications:* <http://dumontierlab.com>

*Presentations:* <http://slideshare.com/micheldumontier>